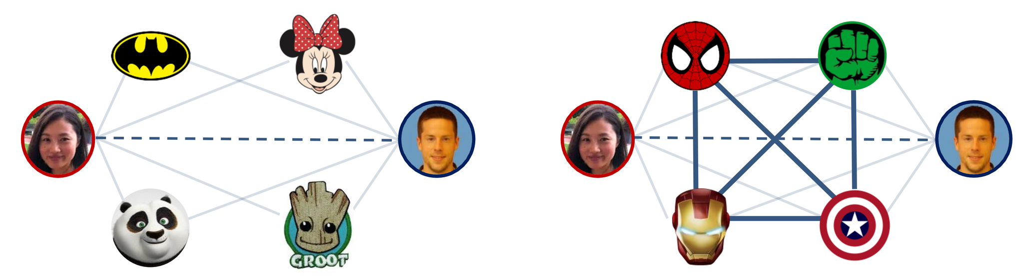
Structural Diversity and Homophily: A Study Across More Than 100 Big Networks

Yuxiao Dong (yuxdong@microsoft.com & ydong1@nd.edu) **Reid A. Johnson, Jian Xu, Nitesh V. Chawla** (rjohns15|jxu5|nchawla@nd.edu)
University of Notre Dame & Microsoft Research   University of Notre Dame

UNIVERSITY OF NOTRE DAME   Microsoft Research

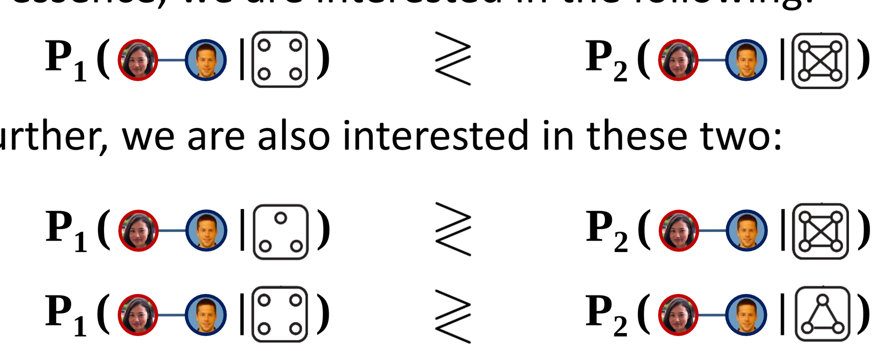
Problem



♣Given that [photo] and [photo] share four common neighbors, are they more likely to connect with each other if their four common neighbors do not know each other (left), or if they all know each other (right)?

♣In essence, we are interested in the following:

♣Further, we are also interested in these two:



Common Neighborhood Signature (CNS)

♣Given a network *G = (V, E)*, its common neighborhood signature is defined as a vector v of relative link existence rates with respect to the specified common neighborhoods. Each element of this vector is a relative link existence rate corresponding to a particular common neighborhood structure.

- *For each network, we get its common neighborhood signature v;*
- *For each pair of networks, we compute the correlation coefficient $\rho(v_i, v_j)$ between their common neighborhood signatures v_i, v_j.*
- *For the similarity matrix, we cluster it hierarchically.*

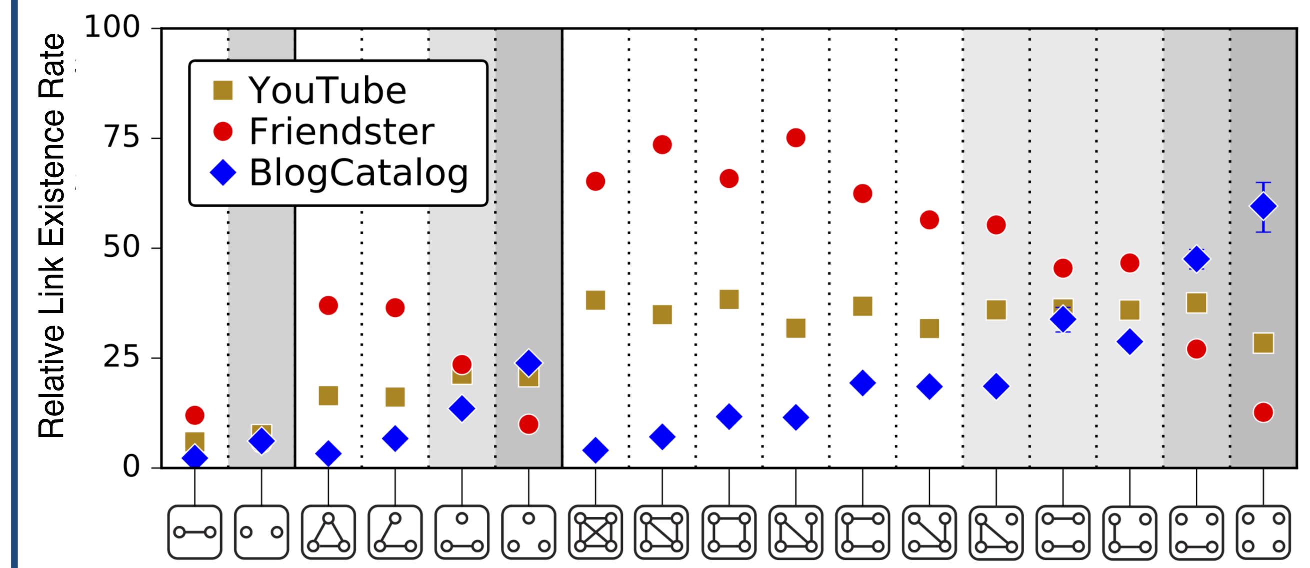
Big Network Data

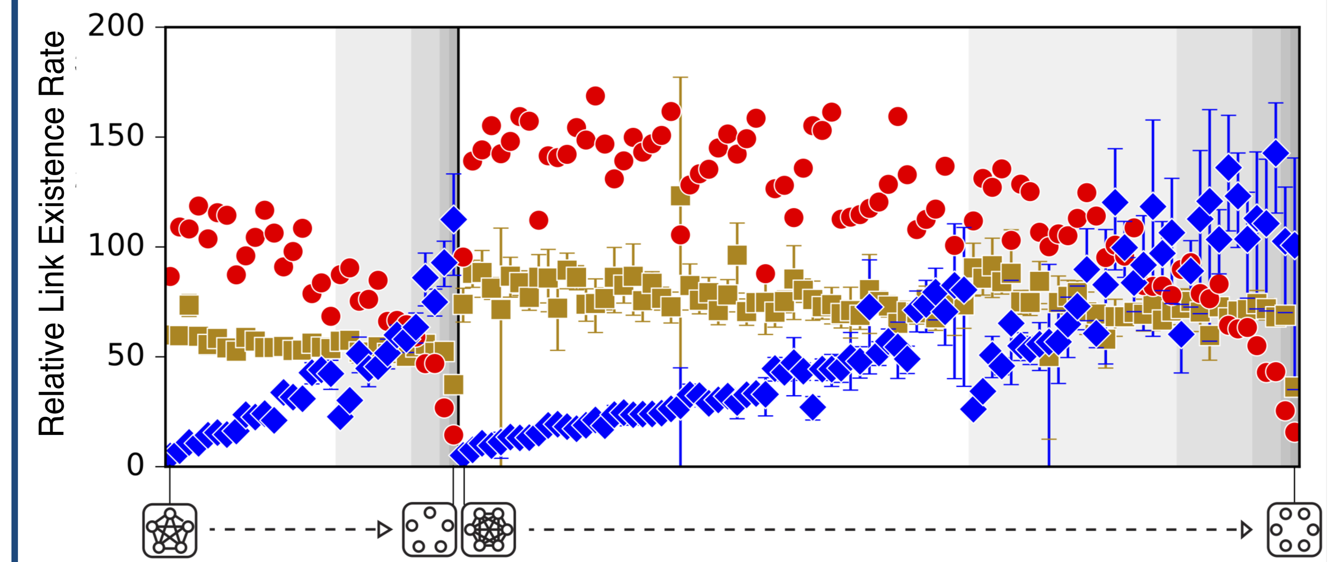
♣80 real networks
- *AMiner.org*
- *ASU*
- *KONECT*
- *MPI-SWS*
- *Notre Dame*
- *Net Repo*
- *Newman*
- *SNAP*

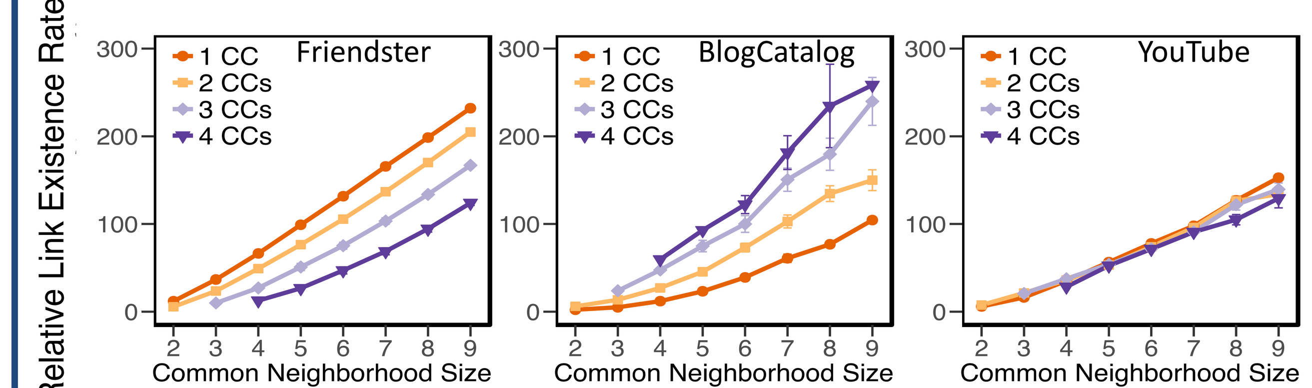
♣40 random graphs by
- *Erdős–Rényi model*
- *Barabási–Albert model*
- *Watts and Strogatz model*
- *Kronecker model*

♣10 for each model with different parameter settings.

Structural Diversity of Common Neighborhoods

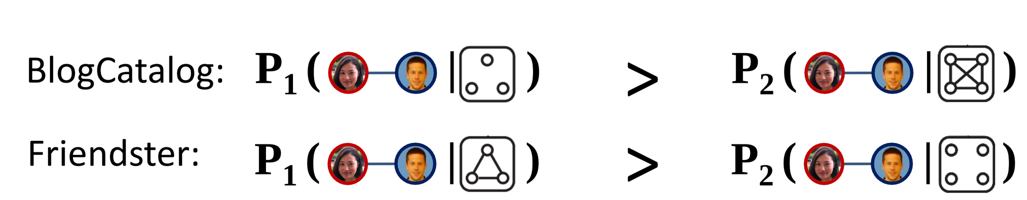






♣The structural diversity of common neighborhoods is a crucial factor in determining link existence across different networks.

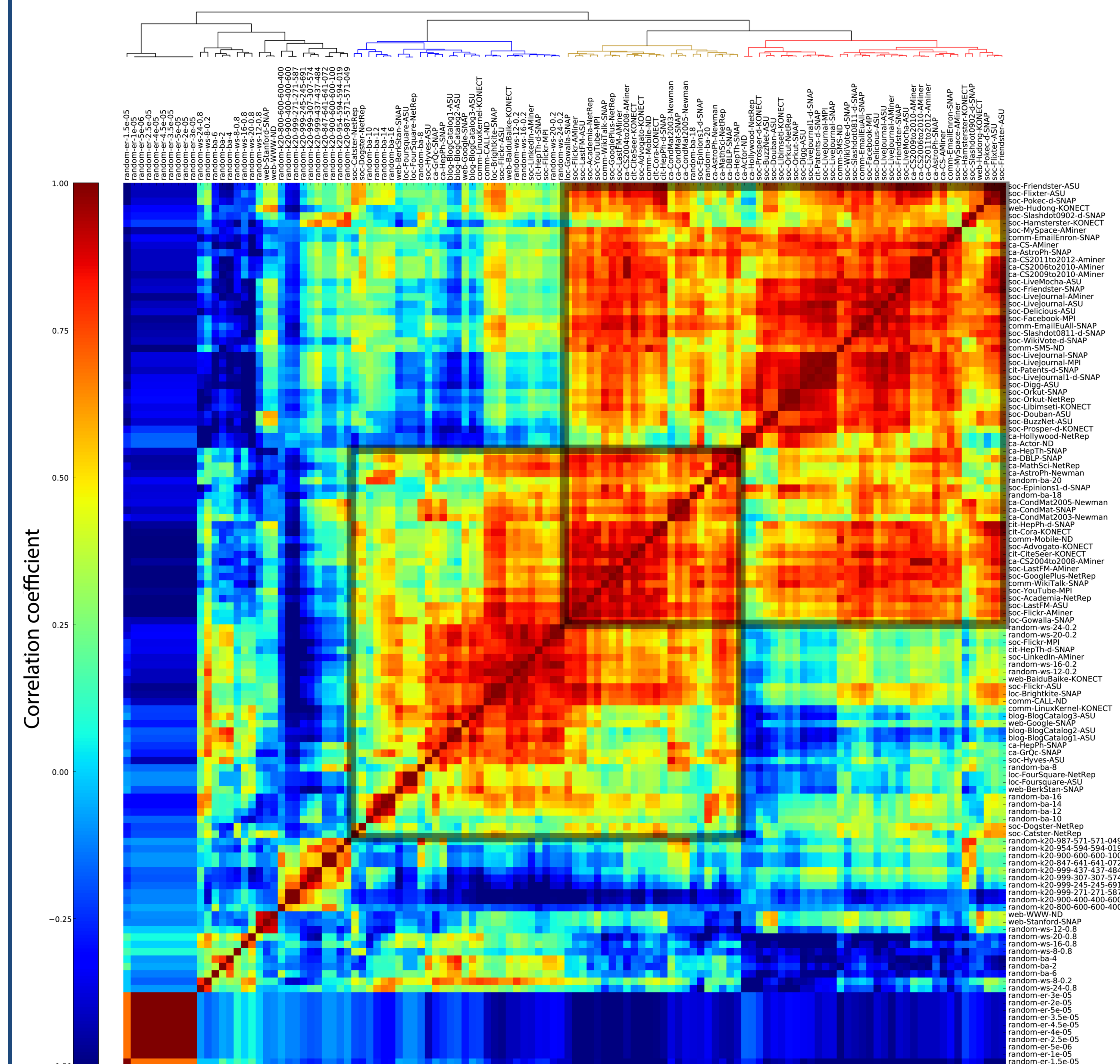
♣When homophily (#CN) is fixed, the structural diversity of common neighborhoods has a negative effect on the formation of online friendships in Friendster but a positive effect in BlogCatalog, and a relatively neutral effect on YouTube.



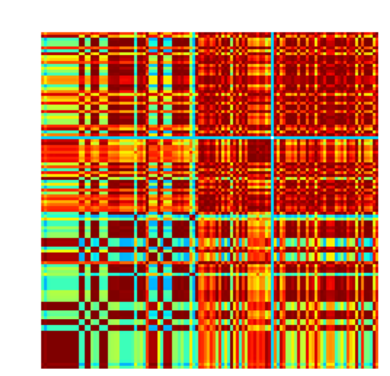
♣Structural diversity, in many cases, violates the principle of homophily, suggesting the fundamental assumption held by the homophily principle can often be an oversimplification.

♣The observations reveal a fundamental difference between these three networks in their microscopic structures and link formation mechanisms.

Network Superfamilies

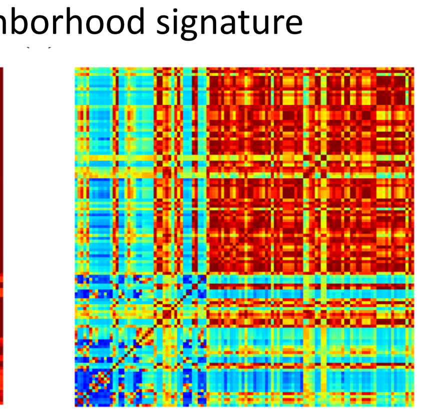


common neighborhood signature

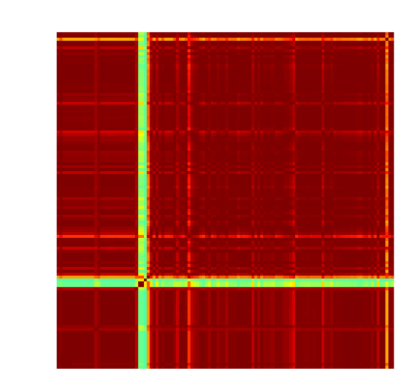


subgraph significance profile

degree sequence



bag of degrees



bag of #common-neighbors

♣Common Neighborhood Structure can detect intrinsic, hidden network superfamilies that are not discoverable by conventional methods.

♣The difference between uncovered superfamilies lie in the distinct strategies that people use across different networking services for satisfying various needs, such the use of Friendster ('red' family) for satisfying social needs and BlogCatalog ('blue' family) for satisfying information needs.

♣Together with classical network properties, we also find that CNS can be used to examine the fitness of random graphs in simulating real networks.

Link Prediction

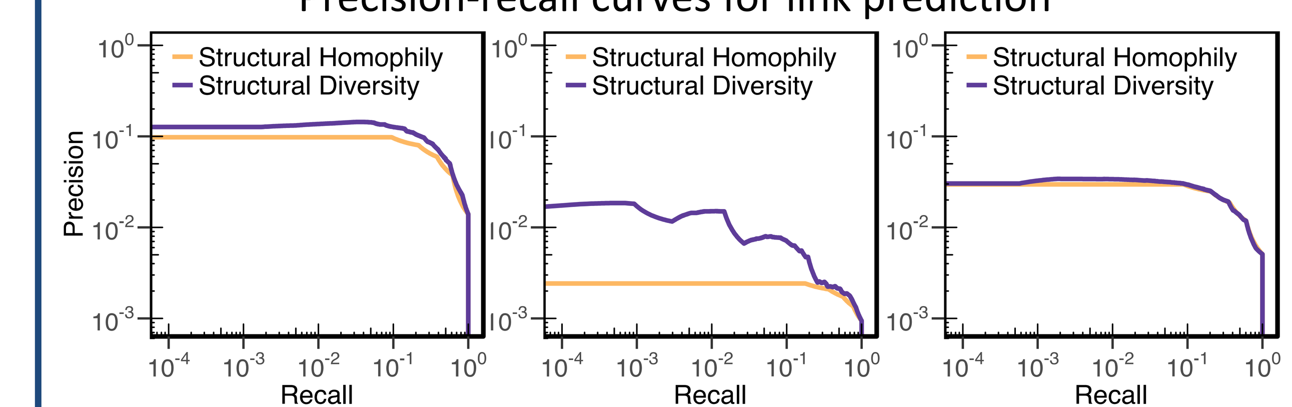
Regression analysis for relative link existence rate

Network	Friendster	BlogCatalog	YouTube
Intercept	−0.03845 ***	0.00010	−0.01855 ***
Homophily (#CN)	0.01948 ***	0.00252 ***	0.00792 ***
Diversity (#components)	−0.01102 ***	0.00114 ***	−0.00047
Adj. R^2 (Diversity)	0.83330	0.76750	0.81440
Adj. R^2 (Homophily)	0.42300	0.14260	0.77160

Link prediction by #CN and structural diversity

Metric	Method	Friendster	BlogCatalog	YouTube
Data	#Pairs	67,033,108,105	224,786,028	118,635,122
	%Positive	0.91830%	0.09430%	0.50820%
AUPR	Homophily	0.02230	0.00178	0.01524
	Diversity	0.03499	0.00279	0.01532
AUROC	Homophily	0.68539	0.66259	0.69371
	Diversity	0.71722	0.70239	0.68401

Precision-recall curves for link prediction



♣Empirical evidence shows that the structural diversity of common neighborhoods helps the link inference task for networks in the 'blue' and 'red' superfamilies

♣Proper application of structural diversity has the potential to substantially improve the predictability of link existence, with important implications for improving recommendation functions employed by social networking sites.

Summary

♣The structural diversity of common neighborhood has significant & distinct effects on link formation and network organization across different networks.

♣Common neighborhood signature can uncover unique network superfamilies, in each of which network structures are formed under certain needs---notably social needs (Friendster & Facebook) and information needs (BlogCatalog & LinkedIn).

♣Common neighborhood signature can serve as a new network property for examining real networks and designing random graph generation models.

References

1. J. Ugander, L. Backstrom, C. Marlow, J. Kleinberg. Structural diversity in social contagion. *PNAS, 2012*
2. B. Uzzi. Social structure &competition in interfirm networks: The paradox of embeddedness. *Administrative Science Q., 1997*.
3. R. Milo, et al. Superfamilies of evolved and designed networks. *Science 2009.*
4. M. Granovetter. *Economic Action and Social Structure: The Problem of Embeddedness. American Journal of Sociology, 1985.*
5. J. Leskovec, L. Backstrom, R. Kumar, A. Tomkins. *Microscopic evolution of social networks. KDD, 2008.*

Acknowledgements

ARL   NSF

Paper Information

KDD2017

Yuxiao Dong, Reid A. Johnson, Jian Xu, Nitesh V. Chawla.
Structural diversity and homophily: A study across more than one hundred big networks.
In *Proc. of 23rd ACM SIGKDD Conf. on Knowledge Discovery and Data Mining* (***KDD'17***).

Data & Code